Vol.2022-SLP-140 No.15
2022/3/1

End-to-End方言音声認識のための方言ラベルを考慮した半教師あり学習

今泉 遼[1,a)]　増村 亮[2,b)]　塩田 さやか[1,c)]　貴家 仁志[1,d)]

概要：本論文では, 方言に対してロバストな End-to-End 音声認識のための方言ラベルを考慮した半教師あり学習を提案する. 最先端の深層学習に基づく音声認識手法である End-to-End 音声認識モデルは学習に大量の音声と書き起こしのペアデータが必要となることが知られている. また, これまでに音声と書き起こしのペアデータのうち書き起こしだけがない音声データが大量にある場合, 音声データをモデル学習に有効活用する方法として半教師あり学習が提案されている. 半教師あり学習法の 1 つとして, ペアデータが揃っている小規模な学習データを用いて教師モデルを学習し, 教師モデルを用いて生成した自動書き起こしをペアデータとして補完することで大規模なデータの活用を可能としている. 日本語方言のための方言ラベルを考慮した End-to-End 音声認識モデルでは書き起こしだけでなく方言ラベルを用いてモデル学習を行なっている. しかし, 音声データに対して方言ラベルと書き起こしのペアデータが揃っているデータの収集は困難である. そこで本研究では, 方言ラベルのみ付与された音声データが大量にある場合を考え, 方言音声データの活用のために方言ラベルを考慮した音声認識モデルのために半教師学習を用いることを提案する. 提案法では, 方言ラベルを考慮することで方言の書き起こしの精度を向上することが可能となるため, 最終的な半教師あり学習によって得られるモデルが方言に対して頑健になることが期待できる. 実験において提案法は方言ラベルを考慮しない半教師あり学習よりも CER が改善することを報告する.

キーワード：End-to-End 音声認識, 多方言音声認識, 半教師あり学習, 方言ラベル

1. はじめに

音声認識 (Automatic speech recognition, ASR) とは話し言葉を文字列に変換する技術である. 近年, 深層学習を用いた手法の 1 つである End-to-End (E2E) ASR モデルが提案され, 高い認識性能が得られることから活発に研究されている [1,2]. 高性能な E2E ASR モデルの構築には大量の音声データとその音声の書き起こしのペアデータが必要となる [3,4]. また, E2E ASR モデルの認識性能は書き起こしの精度に大きく影響することが知られている. しかし, 音声に対応した正確な書き起こしを用意するには人手による書き起こしが必要となり膨大なコストがかかってしまう. 一方, スマートフォンやスマートスピーカーの普及に伴い音声を収録する敷居は低くなり, 音声のみのデータは大量に入手可能となってきている. 書き起こしを含まない大量の音声データを E2E ASR モデルの構築に有効活用するための枠組みの 1 つとして半教師あり学習がある [5–7]. 半教師あり学習手法の 1 つは, 少量の人手による書き起こしと音声のペアデータを用いて教師となる E2E ASR モデルを学習し, その教師モデルを用いて大量の音声データから自動書き起こしを生成し, 人手および自動で付けられた書き起こしを合わせて新たなモデルを学習するものである. 半教師あり学習は ASR だけでなく多言語音声認識や感情認識など様々なデータ収集やラベル付けが困難なタスクで用いられ, 性能が向上することが報告されている [8,9].

日本語の方言も書き起こしが難しく, 音声と書き起こしのペアデータが少ないことが知られている. また, 標準語と比較してアクセントや未知の単語が多いという理由からも入力に方言音声を用いた時, E2E ASR モデルは認識性能が低下することが報告されている [10]. 従来研究では方言識別と音声認識のマルチタスク学習を行うことで認識性能が改善すると報告されている [11]. しかしながら, 従来手法では方言と標準語において大量の音声と人手による正確

[1] 現在, 東京都立大学 システムデザイン研究科 情報科学域
Presently with Tokyo Metropolitan University, Faculty School of Systems Design, Department of Computer Science
[2] 日本電信電話株式会社
Nippon Telegraph and Telephone
[a)] imaizumi-ryo@ed.tmu.ac.jp
[b)] ryou.masumura.ba@hco.ntt.co.jp
[c)] sayaka@tmu.ac.jp
[d)] kiya@tmu.ac.jp

な書き起こしがあることが前提となっており, ペアデータが少ない場合は考慮されていない.

そこで, 本研究では従来手法の 1 つである方言識別の結果を用いて認識を行う E2E ASR モデル (DID2ASR) を用いた半教師あり学習を提案する. まず, 本研究の前提として, 標準語は音声と書き起こしのペアデータが大量にあり, 方言データは音声と書き起こしのペアデータが少量, 書き起こしが付与されていない音声データが大量にあるものとする. ただし, 方言データにはどの地方の方言であるかという方言ラベルは付与されていることとする. この条件のもと書き起こしのついていない大量の方言音声を有効活用するために教師用 DID2ASR モデルを構築する. 用いるデータとしては標準語および少量の方言音声, 人手で書き起こされた書き起こしと方言ラベルとなる. そして, 構築された DID2ASR モデルを用いて大量の方言音声の自動書き起こしを作成して新しいモデル構築を行う半教師あり学習をする. 提案する半教師あり学習では自動で書き起こしを生成する際に方言ラベルを考慮して生成することで信頼性の高い書き起こしを生成し DID2ASR モデルを構築することで認識性能を向上させることが期待できる. 実験では, 6 つの方言からなる自作の音声データベースと標準語音声データベースを用いて, Transformer に基づく E2E ASR モデルおよび DID2ASR モデルそれぞれに対して教師あり学習と半教師あり学習を行い性能を比較した. 実験結果より, 提案した手法は従来の半教師あり学習と比較して方言音声に対する認識性能が向上したことを報告する.

2. 従来法

2.1 End-to-End ASR

本章では, 近年, 高い性能を示している E2E ASR モデルについて説明する. E2E ASR モデルは入力として音響特徴量 $\boldsymbol{X}$ が与えられた時に発話内容の書き起こし $\boldsymbol{W} = \{w_1, ..., w_N\}$ の生成確率を予測するものである. ここで w_n は書き起こしの n 番目のトークン, N は書き起こし内のトークンの数を表す. 自己回帰生成モデルの ASR では, 認識モデルのパラメータセット $\boldsymbol{\Theta}_{\text{asr}}$ を用いて $\boldsymbol{W}$ の生成確率を次のように定義する.

$$P(\boldsymbol{W}|\boldsymbol{X};\boldsymbol{\Theta}_{\text{asr}}) = \prod_{n=1}^{N} P(w_n|\boldsymbol{W}_{1:n-1}, \boldsymbol{X};\boldsymbol{\Theta}_{\text{asr}}) \quad (1)$$

ASR では, 音声発話と書き起こしのペアデータから, 以下の式のようにモデルのパラメータを更新する.

$$\mathcal{D}_{\text{pair}} = \{(\boldsymbol{X}^1, \boldsymbol{W}^1), ..., (\boldsymbol{X}^T, \boldsymbol{W}^T)\} \quad (2)$$

ASR のモデルの目的関数は次のように定義される.

$$\mathcal{L}_{\text{asr}}(\Theta_{\text{asr}}) = -\sum_{t=1}^{T}\sum_{n=1}^{N^t} \log P(w_n^t|\boldsymbol{W}_{1:n-1}^t, \boldsymbol{X}^t;\boldsymbol{\Theta}_{\text{asr}}), \quad (3)$$

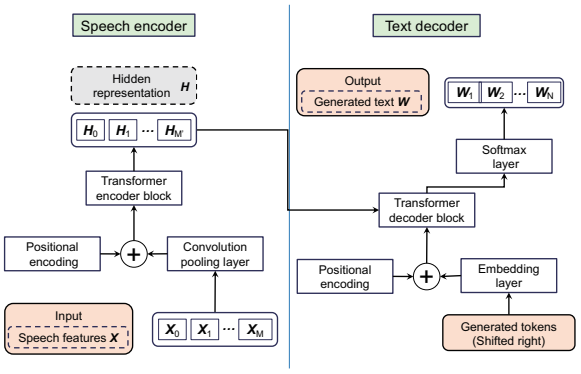

図 1 Transformer に基づく E2E ASR のネットワーク構造

ここで, w_n^t は t 番目の発話の n 番目のトークン, N^t は t 番目の発話のトークンの数を表す.

2.2 Transformer に基づいた End-to-End ASR

本節では, E2E ASR モデルの中でも最先端の性能を示す Transformer に基づいた E2E ASR モデルについて説明する [12–14]. Transformer のネットワーク構造を図 1 に示す. Transformer に基づいた ASR モデルはいくつかの Transformer ブロックで構築される Speech encoder と Text decoder を Sequence-to-sequence モデルとして $P(\boldsymbol{W}|\boldsymbol{X};\boldsymbol{\Theta}_{\text{asr}})$ を計算している. パラメータセット $\boldsymbol{\Theta}_{\text{asr}}$ のモデルは, Speech encoder θ_{enc} と Text decoder θ_{dec} に分割される.

Speech encoder：Speech encoder は, I 個の Transformer encoder block を使用して, 入力音響特徴量系列 $\boldsymbol{X}$ を隠れ表現 $\boldsymbol{H}^{(I)}$ に変換する. i 番目の Transformer encoder block は, 以下の式のように TransformerEncoderBlock 関数を用いて, 下位層の入力 $\boldsymbol{H}^{(i-1)}$ から i 番目隠れた表現 $\boldsymbol{H}^{(i)}$ 構成する.

$$\boldsymbol{H}^{(i)} = \text{TransformerEncoderBlock}(\boldsymbol{H}^{(i-1)};\boldsymbol{\theta}_{\text{enc}}) \quad (4)$$

隠れ表現 $\boldsymbol{H}^{(i)} = \{\boldsymbol{h}_1^{(0)}, ..., \boldsymbol{h}_{M'}^{(0)}\}$ は, 位置情報を埋め込んだ連続ベクトルを追加する AddPostionalEncoding 関数を用いて以下のように定義される.

$$\boldsymbol{h}_{m'}^{(0)} = \text{AddPostionalEncoding}(\boldsymbol{h}_{m'}) \quad (5)$$

$\boldsymbol{H} = \{\boldsymbol{h}_1, ..., \boldsymbol{h}_{M'}\}$ は, 畳み込み層とプーリング層で構成される ConvolutionPooling 関数で定義される.

$$\boldsymbol{H} = \text{ConvolutionPooling}(\boldsymbol{x}_1, ..., \boldsymbol{x}_{M'};\boldsymbol{\theta}_{\text{enc}}) \quad (6)$$

プーリングにより入力 $\boldsymbol{X}$ はサブサンプリングされ, M' はサブサンプリングされたシーケンスの長さを表す.

Text decoder：Text decoder は, 先行するトークンと音声の隠れ表現からトークンの生成確率を計算する. n 番目のトークン w_n の予測確率は, 次のように計算される.

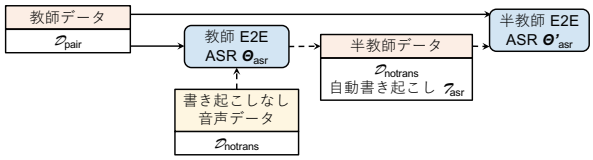

図 2 End-to-End ASR における半教師あり学習

$$P(w_n|\boldsymbol{W}_{1:n-1}, \boldsymbol{X}; \boldsymbol{\Theta}_{\text{asr}}) = \text{Softmax}(\boldsymbol{u}_{n-1}^{(J)}; \boldsymbol{\theta}_{\text{dec}}) \quad (7)$$

ここで, Softmax 関数は線形変換を使用したソフトマックス層を表し $\boldsymbol{u}_{n-1}^{(J)}$ は, J 個の Transformer decoder block から計算される. j 番目の Transformer decoder block は, 以下の式によって, TransformerDecoderBlock 関数を用いて下位の入力 $\boldsymbol{U}_{1:n-1}^{(j-1)} = \{\boldsymbol{u}_1^{(j-1)}, ..., \boldsymbol{u}_{n-1}^{(j-1)}\}$ から j 番目の隠れ表現 $\boldsymbol{u}_{n-1}^{(j)}$ を構成する.

$$\boldsymbol{u}_{n-1}^{(j)} = \text{TransformerDecoderBlock}(\boldsymbol{U}_{1:n-1}^{(j-1)}, \boldsymbol{H}^{(I)}; \boldsymbol{\theta}_{\text{dec}}) \quad (8)$$

隠れ表現 $\boldsymbol{U}_{1:n-1}^{(0)} = \{\boldsymbol{u}_1^{(0)}, ..., \boldsymbol{u}_{n-1}^{(0)}\}$ は, 以下の式で表される.

$$\boldsymbol{u}_{n-1}^{(0)} = \text{AddPostionalEncoding}(\boldsymbol{w}_{n-1}) \quad (9)$$

$$\boldsymbol{w}_{n-1} = \text{Embedding}(w_{n-1}; \boldsymbol{\theta}_{\text{dec}}) \quad (10)$$

Embedding 関数は, 入力トークンを連続ベクトルに埋め込む線形層である.

2.3 End-to-End ASR のための半教師あり学習

本節では E2E ASR モデルにおける半教師あり学習について述べる. 半教師あり学習では式 (2) で示されるデータ $\mathcal{D}_{\text{pair}}$ と以下の式で用いられるデータ $\mathcal{D}_{\text{notrans}}$ を使用する.

$$\mathcal{D}_{\text{notrans}} = \{\boldsymbol{X}^{T+1}, ..., \boldsymbol{X}^{T+L}\} \quad (11)$$

ここで, L は書き起こしなし音声に含まれる発話の数である. 図 2 に半教師あり学習の手順を示す. まず, 2.1 節と同様に教師データ $\mathcal{D}_{\text{pair}}$ を用いてパラメータ $\boldsymbol{\Theta}_{\text{asr}}$ をもつ教師 ASR モデルを構築する. そして, 構築した教師 ASR モデルを用いて, 書き起こしの付与がされていないデータ $\mathcal{D}_{\text{notrans}}$ の認識をし, 自動書き起こし $\boldsymbol{W}$ を生成, $\mathcal{T}_{asr} = \{\boldsymbol{W}^{T+1}, ..., \boldsymbol{W}^{T+L}\}$ となるデータを定義する. 次に, 半教師 ASR モデルは $\mathcal{D}$pair および $\mathcal{D}_{\text{notrans}}$ を用いて学習され, 目的関数は以下のように定義される.

$$\mathcal{L}'_{\text{asr}}(\boldsymbol{\Theta}'_{\text{asr}}) = -\sum_{t=1}^{T+L}\sum_{n=1}^{N^t} \log P(w_n^t|\boldsymbol{W}_{1:n-1}^t, \boldsymbol{X}^t; \boldsymbol{\Theta}'_{\text{asr}}) \quad (12)$$

この半教師 ASR モデルを用いることで, 音声のみのデータを有効に活用し, データ量が増え, 性能の高い認識が可能になると報告されている. しかしながら, 音声データには話者情報や方言ラベルなど書き起こしはないが, 別のラベルが付いている場合がある. その場合, E2E ASR モデルのための半教師あり学習は音声と書き起こしのペアデータを用いて認識するため, 別のラベル情報を活用することができない. また, 教師データと半教師データのドメインや収録条件が異なる場合, 半教師データが性能の向上に寄与せず, 反対に悪影響を及ぼしてしまうこともある.

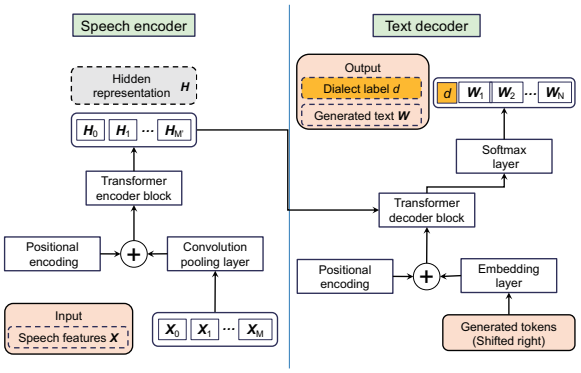

図 3 方言ラベルを考慮した ASR (DID2ASR) のネットワーク構造

3. 提案法

End-to-End 方言音声認識のための方言ラベルを考慮した半教師あり学習の説明をする.

3.1 DID2ASR

方言ラベルを考慮した Transformer に基づく E2E ASR モデルを “DID2ASR” と呼び, このモデルの説明をする [11]. DID2ASR では方言識別を行い, 推定された方言情報を用いて ASR を行う. DID2ASR は 2.1 で示した書き起こしの生成確率の式 $P(\boldsymbol{W}|\boldsymbol{X}; \boldsymbol{\Theta})$ に方言ラベル d を加えたものであり, 書き起こしや方言ラベルの生成確率は次式のように再定義される.

$$\begin{aligned} P(\boldsymbol{W}, d|\boldsymbol{X}; \boldsymbol{\Theta}_{\text{d2a}}) &= P(\boldsymbol{W}|\boldsymbol{X}, d; \boldsymbol{\Theta}_{\text{d2a}}) P(d|\boldsymbol{X}; \boldsymbol{\Theta}_{\text{d2a}}) \\ &= P(\boldsymbol{Z}|\boldsymbol{X}; \boldsymbol{\Theta}_{\text{d2a}}) \end{aligned} \quad (13)$$

$$P(\boldsymbol{Z}|\boldsymbol{X}; \boldsymbol{\Theta}_{\text{d2a}}) = \prod_{n=0}^{N} P(z_n|\boldsymbol{Z}_{1:n-1}, \boldsymbol{X}; \boldsymbol{\Theta}_{\text{d2a}}) \quad (14)$$

ここで, 出力列は方言ラベル d, 書き起こし列 $\boldsymbol{W}$ の順番で連結され, $\boldsymbol{Z} = \{d, w_1, ..., w_N\}$ として定義される. 図 3 より, DID2ASR の Speech encoder は 2.1 節に示す Transformer に基づいた E2E ASR モデルの構成と同様のものであり, Text decoder は, 2.2 節に示した構成に方言ラベルを加え, 方言ラベル d と書き起こし $\boldsymbol{W}$ の両方の生成確率に Softmax 関数を適用する. モデルパラメータは音声, 方言ラベル, 書き起こしのセットを用いて最適化をする.

$$\mathcal{D}_{\text{set}} = \{(\boldsymbol{X}^1, d^1, \boldsymbol{W}^1), ..., (\boldsymbol{X}^T, d^T, \boldsymbol{W}^T)\} \quad (15)$$

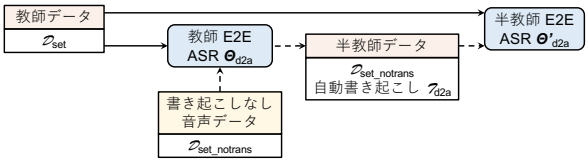

図 4 DID2ASR における書き起こし生成時に方言ラベルを考慮した半教師あり学習 (提案法)

また, 目的関数は $\boldsymbol{Z}$ を用いて以下のように定義される.

$$\mathcal{L}_{\text{d2a}}(\boldsymbol{\Theta}_{\text{d2a}}) = -\sum_{t=1}^{T} \log P(\boldsymbol{W}^t, d^t | \boldsymbol{X}^t; \boldsymbol{\Theta}_{\text{d2a}})$$
$$= -\sum_{t=1}^{T}\sum_{n=1}^{|\boldsymbol{Z}^t|} \log P(z_n^t | \boldsymbol{Z}_{1:n-1}^t, \boldsymbol{X}^t; \boldsymbol{\Theta}_{\text{d2a}}) \quad (16)$$

3.2 方言を考慮した半教師あり学習

DID2ASR を用いて自動で書き起こしを生成する際に方言ラベルを考慮して生成する半教師あり学習を提案する. 方言を考慮した半教師あり学習は式 (15) で示されるデータ $\mathcal{D}_{\text{set}}$ と以下の式で定義されるデータ $\mathcal{D}_{\text{set\_notrans}}$ を使用する.

$$\mathcal{D}_{\text{set\_notrans}} = \{(\boldsymbol{X}^{T+1}, d^{T+1}), ..., (\boldsymbol{X}^{T+L}, d^{T+L})\} \quad (17)$$

ここで, 図 4 に方言を考慮した半教師あり学習の実行手順を示す. 2.3 節と同様に書き起こしなしデータ $\mathcal{D}_{\text{set\_notrans}}$ を認識し, 自動書き起こしと方言ラベルを連結した出力列 $\boldsymbol{Z}$ を生成, $\mathcal{T}_{d2a} = \{\boldsymbol{Z}_{T+1}, ..., \boldsymbol{Z}_{T+L}\}$ となるデータを定義する. 半教師 DID2ASR は $\mathcal{D}_{\text{set}}$ および $\mathcal{D}_{\text{set\_notrans}}$ から学習され, 目的関数を以下のように再定義する.

$$\mathcal{L}'_{\text{d2a}}(\boldsymbol{\Theta}'_{\text{d2a}}) = -\sum_{t=1}^{T} \log P(\boldsymbol{W}^t, d^t | \boldsymbol{X}^t; \boldsymbol{\Theta}'_{\text{d2a}})$$
$$= -\sum_{t=1}^{T}\sum_{n=1}^{|\boldsymbol{Z}^t|} \log P(z_n^t | \boldsymbol{Z}_{1:n-1}^t, \boldsymbol{X}^t; \boldsymbol{\Theta}'_{\text{d2a}}) \quad (18)$$

[11] では, DID2ASR において方言ラベルが既知である場合に認識性能が高いことが示されていることから, 半教師あり学習でも方言ラベルを考慮して書き起こしを生成することで 方言と標準語の情報の混在を防ぎ信頼性の高い書き起こしを生成することが可能である. この信頼性の高い書き起こしを用いることで ASR の性能が向上すると考えられる.

4. 実験

4.1 データベース

本実験で使用するデータベースには自作の日本語方言音声データベース [11] と標準語音声データベースの 2 つを用いた. 方言データベースは, 青森, 広島, 熊本, 名古屋, 札幌, 仙台の 6 地方の方言から構成されており, 標準語データベースには, 日本語話し言葉コーパス (CSJ) [15] を用いた. 表 1 にデータベースの発話数の詳細を示す. 教師データは音声と書き起こしおよび方言ラベルのペアデータ, 半教師データは音声と方言ラベルのペアデータを表す. 教師データと半教師データには発話内容の一部重複があるが, 話者は異なっている. テストデータは教師および半教師と同じ発話内容, 話者は含まない. 方言データにおける方言ごとの男女比は偏りがない. 各方言発話は iPhone5 または XperiaZ1 を用いて収録されており, 日常会話をメインとした 7 秒程度のものとなっている. 標準語の日本語データベースには学術講演と模擬講演からなる CSJ (Corpus of Spontaneous Japanese) を使用した. 男性話者の数は女性話者の約 2 倍であった. 音声データベースの書き起こしおよび DID2ASR において用いられる方言ラベルは人手で付与されている. 全データベースのサンプリング周波数は 16kHz, 量子化ビットは 16bit となっている.

表 1 日本語方言データベースおよび標準語データベースの発話数の内訳

		教師 ($\boldsymbol{X}$, $\boldsymbol{W}$, d)	半教師 ($\boldsymbol{X}$, d)	開発 ($\boldsymbol{X}$, d)	テスト ($\boldsymbol{X}$, d)
方言	青森	1000	9741	676	676
	広島	1000	17670	566	567
	熊本	1000	8328	719	719
	名古屋	1000	17611	551	551
	札幌	1000	14955	678	678
	仙台	1000	15512	535	535
方言全体		6000	83817	3725	3726
標準語		162243	-	1292	2573

4.2 実験条件

提案する方言音声認識の詳細な条件を示す. Speech encoder のエンコーダーブロック数は $I = 8$, Text decoder のデコーダーブロック数は $J = 6$ とした. Transformer ブロックの構成については, 出力連続表現を 256 次元, 位置ごとの Feed forward ネットワークの内部出力を 2,048 次元, Multi-head attention のヘッド数を 4 とした. Speech encoder では入力に音響特徴量として 40 次元のログメルスケールフィルターバンクにデルタおよび加速係数を追加して使用した. フレーム長は 25 ms, フレームシフトは 10 ms とした. また音響特徴量は, ストライドが 2 の 2 つの畳み込み層とマックスプーリング層を通過したため, 時間軸に沿って 1/4 にダウンサンプリングする. Text decoder では, 256 次元の単語埋め込みを使用し, ビームサイズが 20 に設定されたビーム検索アルゴリズムを使用した. ネットワークの最適化には学習率 0.0001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-9}$ の radam オプティマイザーを用いた. ミニバッチサイズは 16, Transformer ブロックのドロップアウト率は

表 2 教師モデルにおける半教師データの CER(%) と ACC(%)

条件	教師モデル	方言ラベル (半教師)	CER	ACC
A	E2E ASR $\boldsymbol{\Theta}_{\text{asr}}$	-	12.0	-
B	DID2ASR $\boldsymbol{\Theta}_{\text{d2a}}$	推定	11.3	53.1
C	DID2ASR $\boldsymbol{\Theta}_{\text{d2a}}$	既知	10.1	100

表 3 教師モデルにおけるテストデータの CER(%) と ACC(%)

教師モデル	方言ラベル (テスト)	CER 方言	CER 標準	ACC
E2E ASR $\boldsymbol{\Theta}_{\text{asr}}$	-	22.9	14.5	-
DID2ASR $\boldsymbol{\Theta}_{\text{d2a}}$	推定	23.4	12.2	63.2
DID2ASR $\boldsymbol{\Theta}_{\text{d2a}}$	既知	22.1	12.2	100

0.1 を設定した. また, データ拡張として SpecAugment, 過学習を防ぐためにラベルスムージングを用いた. 本実験は方言識別の影響も調べるため, 評価指標は次式による文字誤り率 (CER) と方言識別正解率 (ACC) の 2 つを用いた.

$$\text{CER} = (1 - \frac{\text{文字正解率} - \text{挿入語数}}{\text{全文字数}}) \times 100(\%) \quad (19)$$

$$\text{ACC} = (\frac{\text{正解ファイル数}}{\text{全ファイル数}}) \times 100(\%) \quad (20)$$

比較手法として一般的な Transformer に基づく E2E ASR モデルのための半教師あり学習を従来法とした. DID2ASR を用いた半教師学習において方言ラベルを未知として推定する従来法 [11] と本論文の提案法である方言ラベルを既知とした半教師あり学習の 3 手法を用いた.

4.3 実験結果

比較する 3 手法の教師モデルの性能を比較するために半教師用音声データの CER と DID2ASR モデルで推定された方言ラベルの ACC を表 2 に示す. 条件 A は Transformer による E2E ASR $\boldsymbol{\Theta}_{\text{asr}}$ を教師モデルとした場合を表す. DID2ASR $\boldsymbol{\Theta}_{\text{d2a}}$ は方言ラベルを推定または既知として扱える学習モデルである. そのため, 条件 B は推定した方言情報を用いて学習したモデルを表し, 条件 C は方言ラベルが既知, つまり, ACC が 100%の条件で DID2ASR を学習したモデルとする. 表 2 より, DID2ASR モデルは方言ラベルの推定精度によらず方言ラベルを考慮しない E2E モデルよりも CER が低くなっていることから DID2ASR モデルの有効性が確認できる. また, 提案法で想定している DID2ASR のラベルが既知である条件 C では方言ラベルを推定する条件 B の場合よりも CER が改善していることから条件 C の場合, より正確な書き起こしの生成が可能であることが確認できた.

次に, テストデータに対する各教師モデルの CER を表 3 に示す. 教師モデルは学習データにおける方言の割合が低

表 4 半教師モデルにおけるテストデータの CER(%) と ACC(%)

条件	半教師モデル	方言ラベル (半教師)	方言ラベル (テスト)	CER 方言	CER 標準	ACC
A	$\boldsymbol{\Theta}'_{\text{asr}}$	-	-	18.2	16.6	-
B	$\boldsymbol{\Theta}'_{\text{d2a}}$	推定	推定	18.6	13.5	62.3
	$\boldsymbol{\Theta}'_{\text{d2a}}$	推定	既知	17.7	13.5	100
C	$\boldsymbol{\Theta}'_{\text{d2a}}$	既知	推定	18.1	12.9	82.7
	$\boldsymbol{\Theta}'_{\text{d2a}}$	既知	既知	17.5	12.9	100

くテストデータの発話内容が未知であることからどのモデルでも高い CER となっていることがわかる. 特に, 方言ラベルを推定する条件 B の DID2ASR モデルでは方言ラベルの誤識別の影響を受けて方言ラベルを用いない E2E ASR モデルよりも CER が悪化していることがわかる. 一方, 方言ラベルが既知である条件 C の場合は E2E ASR モデルよりも改善していることが確認できた.

表 4 に教師モデルの学習条件 A, B, C それぞれにおいて半教師あり学習を行なった際の CER と推定された方言ラベルの ACC を示す. 表 3 と比較すると, 半教師あり学習を行なった結果どのモデルにおいても方言の CER が大幅に改善していることがわかる. 条件 A の教師モデルにおける CER は 22.9%で半教師モデルにおける CER は 18.2%と大幅に下がっていることから方言音声認識において半教師あり学習は有効であることがわかる. 条件 B ではテスト時に方言を推定した場合, E2E ASR モデルの半教師あり学習と比べて方言推定の誤識別の影響から CER が高くなってしまったが, テスト方言を既知とすることで 17.7%と性能が改善した. 提案法である条件 C においては方言を推定した場合でも半教師データの書き起こしに方言ラベルを考慮したことで方言推定の誤識別の影響が軽減され, CER が 18.1%となり, E2E ASR モデルの CER よりも改善した. 特に, 提案法においてテスト時にも方言が既知である場合, CER は 17.5%と最も低い値を示した. このように提案法は方言ラベルを有効活用可能なモデルであることが示された.

次に, 標準語の CER において表 3 と比較すると, 半教師あり学習を行なった場合に全ての条件で悪化している. この結果は 2.3 節で述べたように, 教師データと半教師データのドメインの違いから半教師データが教師データに含まれる標準語の認識に悪影響を及ぼし, 性能が悪化したと考えられる. しかしながら, 半教師あり学習における提案法の CER は 12.9%, と認識の性能は悪化したが教師モデルとの CER の差が最も小さい. このことより, 半教師データを認識する時に方言ラベルを入力することは半教師あり学習における半教師ありデータが偏ったバイアスをかけて標準語の認識性能を悪化させることを緩和することができる.

5. まとめ

本研究では, End-to-End 方言音声認識のため方言ラベルを考慮した半教師あり学習を提案した. 実験結果より, 方言ラベルを既知として半教師データの書き起こしを生成することで生成される書き起こしの精度が高くなり音声認識性能が改善することがわかった. さらに, テスト時にも方言ラベルを考慮できる場合, 大幅に性能が改善することがわかった. このことから, 方言ラベルを考慮して半教師あり学習を行うことは方言音声認識に対して有効であることがわかった. 今後の課題として, 半教師ありデータの書き起こし生成を動的に行うことや構築された ASR システムを用いてまた書き起こしを生成することを繰り返し, 精度の高い書き起こしを生成することなどがあげられる.

参考文献

[1] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *proc. ICASSP*, pp. 4835–4839, 2017.
[2] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, *et al.*, "Deep speech 2: End-to-end speech recognition in english and Mandarin," in *proc. ICML*, pp. 173–182, 2016.
[3] A. Renduchintala, S. Ding, M. Wiesner, and S. Watanabe, "Multi-modal data augmentation for end-to-end ASR," in *proc. INTERSPEECH*, pp. 2394–2398, 2018.
[4] T. Hayashi, S. Watanabe, Y. Zhang, T. Toda, T. Hori, R. Astudillo, and K. Takeda, "Back-translation-style data augmentation for end-to-end ASR," in *proc. SLT*, pp. 426–433, 2018.
[5] D. S. Park, Y. Zhang, Y. Jia, W. Han, C.-C. Chiu, B. Li, Y. Wu, and Q. V. Le, "Improved noisy student training for automatic speech recognition," in *proc. INTERSPEECH*, pp. 2817–2821, 2020.
[6] G.-A. Levow, "Unsupervised and semi-supervised learning of tone and pitch accent," in *proc. HLT NAACL*, pp. 224–231, 2006.
[7] Y. Higuchi, N. Moritz, J. L. Roux, and T. Hori, "Momentum pseudo-labeling for semi-supervised speech recognition," in *proc. INTERSPEECH*, pp. 726–730, 2021.
[8] J. Cui, B. Kingsbury, B. Ramabhadran, A. Sethy, K. Audhkhasi, X. Cui, E. Kislal, L. Mangu, M. Nussbaum-Thom, M. Picheny, *et al.*, "Multilingual representations for low resource speech recognition and keyword search," in *proc. ASRU*, pp. 259–266, 2015.
[9] Z. Zhang, J. Han, J. Deng, X. Xu, F. Ringeval, and B. Schuller, "Leveraging unlabeled data for emotion recognition with enhanced collaborative semi-supervised learning," *IEEE Access*, vol. 6, pp. 22196–22209, 2018.
[10] R. Imaiuzmi, R. Masumura, S. Shiota, and H. Kiya, "Dialect-aware modeling for end-to-end japanese dialect speech recognition.," in *proc. APSIPA ASC*, pp. 297–301, 2020.
[11] R. Imaizumi, R. Masumura, S. Shiota, and H. Kiya, "End-to-end japanese multi-dialect speech recognition and dialect identification with multi-task learning," *APSIPA Transactions on Signal and Information Processing (accepted)*, 2022.
[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, c. Kaiser, and I. Polosukhin, "Attention is all you need," in *proc. NIPS*, pp. 5998–6008, 2017.
[13] R. Masumura, M. Ihori, A. Takashima, T. Moriya, A. Ando, and Y. Shinohara, "Sequence-level consistency training for semi-supervised end-to-end automatic speech recognition," in *proc. ICASSP*, pp. 7054–7058, 2020.
[14] S. Karita, N. Enrique Yalta Soplin, S. Watanabe, M. Delcroix, A. Ogawa, and T. Nakatani, "Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration," in *proc. INTERSPEECH*, pp. 1408–1412, 2019.
[15] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese.," in *proc. LREC*, pp. 947–952, 2000.